
# AFRL-RH-WP-TR-2011-0081

# GENOME-WIDE ASSOCIATION MAPPING FOR INTELLIGENCE IN MILITARY WORKING DOGS:

## Development of Advanced Classification Algorithm for Genome-Wide Single Nucleotide Polymorphism (SNP) Data Analysis

**Victor T. Chan**
**Camilla A. Mauzy**
**Armando Soto**
**Jessica A. Wagner**
**Biosciences and Protection Division**
**Applied Biotechnology Branch**

**Amy D. Walters**
**Jeanette S. Frey**
**Tiffany M. Hill**
**Henry M. Jackson Foundation**
**For the Advancement of Military Medicine**
**2729 R Street**
**Wright-Patterson AFB OH 45433-5707**

**Karen L. Overall**
**Penn Med Translation Research Laboratory**
**125 S. 31th St.**
**Philadelphia PA 19104-7051**

**Richard M. Wolfe**
**Lonnie R. Welch**
**School of Electrical Engineering & Computer Science**
**329 Stocker Center**
**Ohio University**
**Athens OH 45701-2979**

**Interim Report April 2011**



**Air Force Research Laboratory**
**711th Human Performance Wing**
**Human Effectiveness Directorate**
**Biosciences and Performance Division**
**Applied Biotechnology Branch**
**WPAFB, OH 45433-5707**

# NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

**AFRL-RH-WP-TR-2011-0081**

THIS REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

| | |
|---|---|
| //SIGNED// | //SIGNED// |
| CAMILLA A. MAUZY, Work Unit Manager | F. WESLEY BAUMGARDNER, PhD |
| Applied Biotechnology Branch | Biosciences and Performance Division |
| | Human Effectiveness Directorate |

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government‟s approval or disapproval of its ideas or findings.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| April 2011 | Interim Report | April 2009 – April 2011 |

**4. TITLE AND SUBTITLE**
GENOME-WIDE ASSOCIATION MAPPING FOR INTELLIGENCE IN MILITARY WORKING DOGS: Development of Advanced Classification Algorithm for Genome-Wide Single Nucleotide Polymorphism (SNP) Data Analysis

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
NA

**5c. PROGRAM ELEMENT NUMBER**
62202F

**6. AUTHOR(S)**
*Victor T. Chan; Camilla A. Mauzy; Armando Soto; Jessica A. Wagner;**Amy D. Walters; Jeanette S. Frey; Tiffany M. Hill; ***Karen L. Overall; ****Richard M. Wolfe; Lonnie R. Welch

**5d. PROJECT NUMBER**
ODA

**5e. TASK NUMBER**
WP

**5f. WORK UNIT NUMBER**
ODAWP001

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
**Henry M. Jackson Foundation for the Advancement of Military Medicine
***Penn Med Translation Research Laboratory, 125 S. 31st St, Philadelphia PA 19104-7051 ****School of Electrical Engineering & Computer Science, 329 Stocker Center, Ohio University, Athens OH 45701-2979

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Air Force Materiel Command*
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Biosciences and Performance Division
Applied Biotechnology Branch
Wright-Patterson AFB OH 45433-5707

**10. SPONSOR/MONITOR'S ACRONYM(S)**
711 HPW/RHPB

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
AFRL-RH-WP-TR-2001-0081

**12. DISTRIBUTION / AVAILABILITY STATEMENT**
Distribution A: Approved for public release; distribution unlimited.

**13. SUPPLEMENTARY NOTES** 88ABW-2011-4088, cleared 25 July 2011.

**14. ABSTRACT**
This project collected data to genetically map superior intelligence in the military working dog. A behavioral testing regimen was developed by canine cognitive expert Dr Karen Overall (UPENN) which enabled quantitative intelligence testing of individual dogs and blood samples were taken, and genome-wide SNP typing completed by means of the Affymetrix Canine SNP (single nucleotide polymorphism) Array v2. In order to identify SNP markers for mapping of small-effect-sized genes that contribute to highly complex polygenic traits, it is necessary to develop a more robust computational method for the analysis of SNP profile data. To accomplish this, we are undertaking two parallel efforts, Biologically Guided Feature Selection and Computational Based Feature Synthesis and Classification. As a proof-of-concept, we conducted a classification analysis focused on a subset of tested canines consisting of German Shepherds, Labrador Retrievers, and Belgian Malinois. Using this new classification technique, samples from the three breeds clustered into the correct breed with an accuracy ranging from 89 – 100 %. Classification accuracy was not significantly affected by data process methods (including data cleanup methods) or SNP annotation quality, thus suggesting that this algorithm is highly robust. With further refinement and optimization, this technique could be used to classify complex phenotypes in an unsupervised manner and allow identification of associated SNP markers.

**15. SUBJECT TERMS**
Military working dog genome-wide association study genetic marker intelligence

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Camilla Mauzy |
| U | U | U | SAR | 21 | 19b. TELEPHONE NUMBER *(include area code)* NA |

**Standard Form 298 (Rev. 8-98)**
**Prescribed by ANSI Std. 239.18**

**THIS PAGE INTENTIONALLY LEFT BLANK.**

## TABLE OF CONTENTS

## LIST OF TABLES

**Table** **Page**

## PREFACE

This research was accomplished at the Applied Biotechnology Branch, Human Effectiveness Directorate of the 711th Human Performance Wing (711 HPW/RHPB) of the Air Force Research Laboratory, Wright-Patterson AFB, OH under Dr. John J. Schlager, Branch Chief. This technical report was written for AFRL Work Unit ODAWP001. This project was partially funded by DARPA (in conjunction with UES contract FA8650-08-C-6832). Henry employees were working under Cooperative Agreement with the Henry M. Jackson Foundation, FA8650-05-2-6518.

Subcontracted research on dog sampling was performed with Dr. Karen Overall, University of Pennsylvania, under UES contract FA8650-08-C-6832.

All studies involving animals were approved by the Wright-Patterson Institutional Animal Care and Use Committee, and were conducted in a facility accredited by the Association for the Assessment and Accreditation of Laboratory Animal Care, International, in accordance with the *Guide for the Care and Use of Laboratory Animals*, National Research Council (1996). Studies were conducted under approved Air Force Research Laboratory Institutional Animal Care and Use Committee Protocol AFDR-2009-002A "*Genome-wide Association Mapping for Superior Intelligence in Military Working Dogs*" (Univ. of PA Protocol #802551).

## ACKNOWLEDGEMENTS

The authors would like to thank Dr. Soraya Juarbe-Diaz and Donna Dyer for the behavioral testing and blood collection of the canines.

## SUMMARY

In a collaborative effort between the Air Force Research Laboratory, Human Effectiveness Directorate, Applied Biotechnology Branch (711 HPW/RHPB), and the University of Pennsylvania, this project collected preliminary data to genetically map superior intelligence in the military working dog (MWD). A behavioral testing regimen was developed by canine cognitive expert Dr. Karen Overall which enabled quantitative intelligence testing of individual dogs using scoring based on latency to response time, focused behavior, success-in-effort time, and handedness (to be reported in a later technical report). Behavior testing and DNA collection was conducted on a cohort using canines from United States working dog contractors. Dogs were tested using the Canine Intelligence Testing Protocol (CITP) and a blood sample collected from each animal. Genomic DNA was prepared from the whole blood, and the isolated DNA was subjected to the genome-wide SNP (single nucleotide polymorphism) typing by means of the Affymetrix Canine SNP Array v2. In order to identify SNP markers for the mapping of small-effect-sized genes that contribute to highly complex polygenic traits (such as intelligence), it is necessary to develop a more robust computational method for the analysis of genome-wide SNP profile data. As a proof-of-concept, we conducted a classification analysis, focused on a subset (117) of tested canines consisting of German Shepherds, Labrador Retrievers, and Belgian Malinois. Using this new classification technique, samples from the three breeds clustered into the correct breed with an accuracy ranging from 89 – 100%. Classification accuracy, however, was not significantly affected by data process methods (including data cleanup methods) or SNP annotation quality, thus suggesting that this algorithm is highly robust. With further refinement and optimization, we anticipate that this technique could be used to classify these canine subjects according to their intelligence in an unsupervised manner and identification of the SNP markers responsible for such classification.



**Keywords:** military working dog, genome-wide association study, genetic marker, intelligence, classification technique, clustering analysis

Technical report April 2011

# 1. INTRODUCTION

## 1.1 Canine Whole Genome Mapping

The completion of the canine genome sequence has resulted in many new genetic markers and thus provided unprecedented opportunities for the identification of genes involved in complex polygenic traits (Ostrander, 2000). The whole-genome scanning approach has many attractive aspects, such as the global assessment of linkage disequilibrium (LD) strength and high resolution for mapping the location of trait-associated loci (Amos 2007; Farrall *et al*. 2005; Pearson *et al.* 2008). Although there are multiple sources of genetic variation in mammalian genomes, single nucleotide polymorphisms (SNPs) have emerged as the marker of choice for whole genome linkage and association studies due to their high abundance, stability, and relative ease of scoring (Ding *et al*. 2009). These attributes make whole-genome SNP typing a powerful technique for conducting genome-wide association studies (GWAS). Most of the SNPs used in GWAS are mapping markers, rather than functional mutations (i.e. not a causative mutation). Despite this, a GWAS with an adequate genomic coverage will allow the identification of a subset of these SNPs that may be very close, in term of chromosomal distance, to a functional quantitative trait locus (QTL). The discovery of a SNP associated with the QTL can thus result in an indirect association between the SNP and the trait itself (Sham *et al*. 2009; Almasy, *et al*. 2009). Therefore, association studies based on the underlying principle of LD are significantly facilitated by the whole-genome SNP screening.

The initial Canine Genome Sequencing Project produced a high-quality draft of the genomic sequence of a female boxer (Lindblad-Toh, *et al.* 2005). By comparing this genome sequence with that of other breeds, the project successfully compiled a comprehensive set of SNPs applicable to all dog breeds (Wayne, *et al.* 2007, Ostrander, *et al.* 2005). These selected SNP markers are spaced 25,000 to 30,000 base pairs (bp) apart (in average) and, while not as dense as the human SNP array (averaging 3,000 bp in distance), are nonetheless useful tools for mapping the trait-associated loci of interest (Karlsson, *et al.* 2007). High-throughput analysis of genome-wide SNP markers in the canine genome can now be achieved using scans of commercially available SNP microarrays (Butcher *et al.* 2008, Ostrander *et al.* 2005). Two versions of the canine SNP arrays exist and they provide a different level of genomic coverage. Version 1 has ~27,000 high quality SNPs, while version 2 contains ~50,000 high-quality SNPs (among a total of 127,132 SNPs per chip). Because of the increased coverage, Version 2 was used in this study. This array is a 5-μm format, perfect match (PM) probe only (20 probes/SNP) Whole Genome Sampling Assay (WGSA) Design and contains probe sets for a total of ~127K SNPs. These SNPs were chosen from the map of over 2.5 million SNPs generated as part of the canine genome project and include the majority of the gold set of the Version 1 array (i.e. 26,625 SNPs derived from a panel of 10 diverse breeds). Similarly, a "platinum" set of 49,633 SNPs has been identified using a panel of 10 diverse breeds in the Version 2 array.

Two different library files can be used with the Version 2 arrays. While the library file ***DogSty06m520431*** will show the results for the full set of the SNPs on the chip (i.e. 127,132 SNPs), the library file ***DogSty06m520431P*** will mask out the SNPs that are not included in the "platinum" set and thus only show the results for the 49,633 SNPs that are considered as high-quality. Despite the concern of their annotation quality, some of the SNPs not included in the "platinum" set may in fact be associated with intelligence. Therefore, both library files were used separately to generate two datasets that were analyzed individually.

One of the factors affecting the power of a genetic study is the information content that can be extracted from the samples. While the physical distance between the QTL and SNP markers is not the only factor that influences the strength of LD, it is still considered a main factor in most cases (Borecki *et al.* 2008, Gu *et al.* 1996). Some analyses suggest that a highly dense map with about 500,000 SNP markers spanning the whole genome may be needed for whole-genome association study, while others have shown that strong LD can be extended up to 1 *cM (*centiMorgan) (Gu and Rao, 2003) and thus ~30,000 SNPs will probably be enough for a genome-wide scan. As the Version 2 of the canine SNP array can provide information content for 50-127K SNPs (depending on the library files used in data processing), genome-wide coverage can thus be adequately achieved using the current canine array design.

### 1.2 Computationally Based Feature Synthesis and Classification Algorithm

To identify and group small-effect-sized QTLs contributing to complex polygenic traits, techniques based on feature synthesis and genetic algorithm were explored. Initially, low dimensional feature vectors were synthesized from the original genotyping dataset that has high dimensional feature vectors using co-evolutionary genetic programming (CGP). The synthesized features were obtained by applying a series of operators (composite operator vectors) to the original features. These operators are binary trees with simple operators as the inner nodes and the original features as the leaf nodes. First, the internal nodes of the tree representing the composite operator were randomly determined in a recursive manner. After all the internal nodes are generated, the original features were randomly picked and attached to the leaf nodes. The genetic programming operations were then applied to the binary trees in the order of crossover, mutation and selection. In addition, an elitism replacement method was adopted to keep the best composite operator, in terms of classification accuracy, from generation to generation.

The classification accuracy of a Bayesian classifier in the low dimensional synthesized feature space was used to assess the fitness of the synthesized features, as assessed by classification accuracy. The best-fitted synthesized features were generated using the CGP algorithm through the iteration of the mutation-selection process. To train the algorithm, CGP were used to run the

training data and evolve through the mutation-selection process to select the best composite operator based on the Bayesian classifier in the synthesized feature space. In the testing phase, the synthesized features were generated by applying the composite operator vector to the original features of the testing samples, and the Bayesian classifier used for the classification of the test samples.

As the first step of the development of this methodology, we analyzed the genome-wide SNP profiles of 117 dogs from three breeds (German Shepherd Dog, Belgian Malinois, and Labrador Retriever) using this approach. We were able to classify these dogs into three groups, one for each breed, with 89 – 100% accuracy. The high degree of accuracy of this classification technique in clustering these test subjects into their corresponding breeds in an unsupervised manner strongly suggests that this algorithm may be further developed and optimized for the analysis of complex traits such as intelligence.

## 1. MATERIALS AND METHODS

### 2.1 Animal Testing Procedures

Canines, already working or in advanced training, were tested in this preliminary study. These dogs were owned by three private contractor facilities. Each animal was tested using the Canine Intelligence Test Protocol (CITP) consisting of tests for: Attentiveness, Novelty, Interest, Signaling/showing, Observational learning/showing, and Problem solving/boldness. The behavioral testing data is not part of the analysis of this report. A more in-depth description of the behavioral testing techniques and the data/result of these tests will be described in a separate technical report.

### 2.2 Blood Sample Collection and Genomic DNA Isolation

After behavioral testing, a blood sample (5-6 ml) was collected from each phenotype tested canine via venipuncture of the cephalic vein by a licensed veterinarian. High-molecular-weight genomic DNA was extracted from blood leukocytes using the Qiagen QIAampR DNA Blood Midi Kit as recommended by the manufacturer. Briefly, blood samples were added to the QIAGEN Protease in a 15-ml centrifuge tube. Lysis buffer (AL) was then added to each samples, followed by vigorous shaking for at least 1 minute. The mixture was then incubated at 70 °C for 10 minutes. Ethanol (100%) was added to each sample, followed by vigorous shaking. One half of the supernatant of each sample was then added onto a QIAamp Midi column (placed in a 15 ml centrifuge tube), and the samples centrifuged at 1,850 x g for 3 minutes. After the removal of the filtrate, the remaining half of the supernatant samples was loaded onto a QIAamp Midi column and the centrifugation step was repeated. The bound DNA was washed using the washing buffers AW1 and AW2. High-molecular weight genomic DNA was then recovered

using the elution buffer AE. The purified DNA samples were stored (in small aliquots) at –20º C before being processed for target preparation.

### 2.3 Target Preparation, Chip Hybridization and Detection

The genomic DNA samples were first diluted to 50 ng/µL using the reduced EDTA-TE buffer in a 96-well reaction plate. Restriction digestion of the DNA samples (using Sty I) was initiated by the addition of 14.75 µL Digestion Master Mix to each sample to produce a final volume of 20 µL containing 250 ng genomic DNA, 2 µg BSA and 1 unit Sty I in 1x restriction digestion buffer (NE Buffer #3: 50 mM Tris-HCl, 100 mM NaCl, 10 mM $MgCl_2$ and 1 mM dithiothreitol). The reaction mix was incubated at 37º C for 2 hours in a thermal cycler. Once the digestion was completed, the enzyme was inactivated by heating at 65º C for 20 minutes. Ligation was initiated by the addition of ligation mix containing DNA ligase and the Sty adaptors to the digested DNA samples. After incubating at 16º C for 3 hours, the reaction mix was heated to 70º C for 20 minutes to inactivate the DNA ligase. The ligation products were then diluted 4-fold in AccuGENE® water (Affymetrix) to yield a final volume of 100 µL.

A 10 µL aliquot of the ligation product from each sample was transferred to the corresponding well of a 96-well reaction plate, followed by the addition of the PCR Master Mix (90 µL/sample) to produce a final volume of 100 µL containing 0.1 mmol GC-Melt, dNTPs (0.035 µmol each), 0.45 nmol PCR Primer #002 and 2 µL Titanium Taq DNA Polymerase (50x stock) in 1x Titanium Taq Buffer. PCR was carried out using the following setting:

1. 94º C for 3 minutes (1 cycle)
2. 94º C for 30 sec → 60º C for 45 sec → 68º C for 15 sec (30 cycles)
3. 68º C for 7 minutes (1 cycle)
4. 4º C → HOLD

After the PCR was completed, the reaction plate was centrifuged at 2,000 rpm for 30 seconds to recover the condensates. The PCR products (3µL/sample) were analyzed using gel electrophoresis (2% agarose in TBE buffer). The fragment size of the PCR products ranged from 250 – 1,100 bp.

The PCR products were purified using the Clontech Clean-Up Plate according to the procedure recommended by the manufacturer with three washes using AccuGENE® water, followed by the elution of the PCR products using RB Buffer. The concentration of the purified PCR products was determined by $OD_{260}$ (optical density). Three dilutions for each PCR product were made and quantified independently. The average of the OD measurements for each sample was calculated and used as the final concentration. Once the concentrations of the samples were determined, they were diluted to 2 µg/µL in RB Buffer.

The purified, normalized PCR products were treated with Fragmentation Reagent at 37º C for 35 minutes, followed by heating at 95º C for 15 minutes. The size of the fragmented PCR products was determined using gel electrophoresis (4% agarose in TBE buffer), which indicated that the average fragment size was less than 180 bp. The fragmented targets were labeled using the GeneChip® DNA Labeling Reagent (from Affymetrix) according to the Affymetrix Human Mapping 500K Array Technical Manual. Briefly, 19.5 µL of Labeling Master Mix was added to each sample, and the reaction mix was incubated at 37º C for 4 hours, followed by incubation at 95º C for 15 minutes. The labeled target for each sample was first mixed with 190 µL of hybridization master mix, and the resulting mix was denatured at 95º C for 10 minutes and kept at 49º C until use. The denatured target was then loaded onto a Canine SNP Array v2. The arrays (with hybridization cocktail loaded) were placed into a preheated hybridization oven and allowed to hybridize at 49º C for 18 hours.

After hybridization, the hybridization cocktail was removed from each chip and transferred to a tube. Array Holding Buffer was then added to each array. The washing, staining, and scanning of the hybridized arrays were performed using the Affymetrix Fluidics Station 450 and the GeneChip® Scanner 3000 7G following the Affymetrix Human Mapping 500K Array Technical Manual.

### 2.4 Canine SNP Array Data Processing

Data processing was performed using the snp5 command line software downloaded from Affymetrix to make the genotype calls. Initially, a QC analysis was performed to assess the data quality. The information in the Intensity QC Table indicated the overall performance of the chip analysis. When all steps of the assay are working as expected, the QC call rate is typically >75% for the entire collection of 127K SNPs and >85% for the "platinum" set of SNPs. As described in section 1.1, both library files (***DogSty06m520431*** and ***DogSty06m520431P***) were used so that two datasets consisting of 127K SNPs or 50K "platinum" SNPs were generated for downstream data analysis. Initially, Dynamic Model algorithm was used to perform QC analysis on individual arrays. Once completed, genotype calls of the SNPs were determined using the Bayesian Robust Linear Model with Mahalanobis distance classifier (BRLMM) Algorithm batch analysis tool (Miclaus *et al.* 2010, Hong *et al.* 2010, Hoggart *et al.* 2003).

The 117 canine subjects were genotyped using the Affymetrix canine SNP array 2.0 in three batches. The SNP array datasets were processed using two different approaches:

1. DP Method 1: Each SNP array dataset was processed separately to generate the genotype calls, and the processed datasets were combined into a single large dataset.

2. DP Method 2: The three SNP array datasets were combined into one large dataset, and the resultant dataset was processed to generate the genotype calls.

## 2.5 Unsupervised Breed Assignment Clustering Analysis

### 2.5.1 Clustering Analysis Steps.

The clustering analysis pipeline consists of the following four steps:

1. Data cleanup
2. Creation of a distance matrix
3. Construction of clusters based on the distance matrix
4. Merging clusters based on the distance matrix.

### 2.5.2 Data Cleanup.

To ensure data quality, a three-step filtering process was developed to filter out low-quality SNPs (and samples) prior to downstream data analysis (Lander *et al.* 1995). In the first filter, samples with an overall call rate of <75% were excluded from the dataset. The filtered sample set was then subjected to the second data filter. Any SNP with <90% call rate across all the samples was eliminated from subsequent data analysis. Following these two filtering steps, the final call rate of the remaining samples/SNPs was examined, and samples with a call rate <95% was excluded from the dataset. We reasoned that this data cleanup procedure was especially important when the full set of 127K SNPs dataset was used since some SNPs in this full set are expected to be of suboptimal quality.

Initially we implemented two simple methods to handle missing data (no calls): 1.) removed all SNPs with any missing data points – this filter resulted in the removal of ~80% of the SNPs; and 2.) no data cleanup – the data was coded so that the metric for comparing how the two SNPs are related can account for the missing data. It was decided that if this simple "all or none" approach failed to generate acceptable clustering results, the more sophisticated 3-step data cleanup procedure (as described above) would be implemented.

These datasets, after data cleanup, were then used as input data for the development and validation of the advanced clustering techniques. The primary goal of the analysis was to develop a clustering technique that can separate dogs by breed, solely based on two pieces of information, the SNP profiles and three canine breeds in the population. Neither the information concerning the number of dogs in each breed nor information on any breed-specific SNPs was used as input data. The secondary goal was to evaluate how data processing, data cleanup and SNP annotation quality may affect the final analysis result.

### 2.5.3 Creation of Distance Matrix.

The distance matrix was generated using the following steps:

1. Compare the genotype of each SNP of all sample pairs and numerically code the distance of each pair-wise comparison
   a. Distance = 0, if both alleles are the same
   b. Distance = 1, if only one allele is the same (i.e. if the genotype of a subject is AA or BB and that of the other subject is AB)
   c. Distance = 2, if no allele is the same (i.e. if the genotype of a subject is AA and that of the other subject is BB)
   d. Distance = N/A, if there is a no call (i.e. missing data) in one sample (or in both samples).
2. Summarize the distance of all pair-wise comparison for all samples.

An example of a distance matrix is shown in Table 1. Since this genome-wide SNP typing dataset is not ready for public release at this point, a hypothetical matrix, rather than the actual matrix, is shown.

**Table 1: Distance Matrix**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ……//…… | 117 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 2.33E+05 | 2.09E+05 | 1.77E+05 | 1.25E+05 | 2.41E+05 | 9.55E+04 | 8.75E+04 | 2.35E+05 | 1.12E+05 | …… | 1.37E+05 |
| 2 | | | 1.05E+05 | 8.85E+04 | 2.19E+05 | 2.06E+05 | 1.13E+05 | 2.75E+04 | 2.31E+05 | 1.85E+04 | …… | 7.76E+03 |
| 3 | | | | 9.65E+04 | 7.15E+04 | 9.46E+03 | 2.13E+05 | 7.14E+04 | 1.70E+05 | 3.46E+03 | …… | 3.15E+04 |
| 4 | | | | | 5.55E+04 | 1.19E+05 | 9.25E+04 | 1.91E+05 | 9.95E+04 | 1.69E+05 | …… | 1.15E+05 |
| 5 | | | | | | 1.47E+05 | 1.45E+05 | 7.75E+04 | 5.05E+04 | 1.30E+05 | …… | 1.99E+05 |
| // | | | | | | | | | | | …… | |
| 117 | | | | | | | | | | | | |

**2.5.4 Clustering Algorithm.**

The algorithm used for unsupervised breed assignment analysis was based on the hierarchical clustering technique using the Ward's algorithm for the calculation of the distance-based group assignment (Ward, *et al.* 1961). Specifically, the algorithm and parameters are shown below:

$$d(k.ij)=\{(C_k+C_i)D_{ki}+C_j+C_k)D_{jk}-C_k*D_{ij})\}/C_k+C_i+C_j)$$

where d(k.ij) = the distance between new clusters, $C_{i,j,k}$ = the number of cells in cluster i,j, or k, and $D_{ki}$ = the distance between cluster k and i at the previous stage/iteration.

The analysis starts with 117 clusters, each cluster containing only one sample. The algorithm then identifies the closest pair of clusters and merges them into one single cluster. The distances

between the new cluster and all other clusters are then re-calculated, and the closest pair of clusters identified and merged. This process is reiterated until all the samples are merged in one single cluster. The distance from the root is selected to result in three separate clusters.

## 3. RESULTS

In this study, a total of 117 canine subjects, identified during the behavioral testing phase as German Shepherd Dog (47 dogs), Belgian Malinois (44 dogs) and Labrador Retriever (26 dogs) were selected and genotyped using the Affymetrix canine SNP Array v2 in three batches. The SNP array datasets generated were processed using two different approaches (for details, see Materials and Methods Section 2.4). This finalized dataset, regardless the data processing methods used, thus contained the genotype calls of 127,132 SNPs (distributed across the entire canine genome) of 117 dogs belonging to three breeds. Additionally, a dataset containing the genotype calls of a subset of these SNPs that represent the high-quality SNP set (49,663 SNPs) was also generated using the Platinum Set library file. These datasets were then used as the input files for the analyses. As mentioned above, we were interested in developing an advanced clustering technique that can separate the dogs by breed, solely based on the two pieces of information, i.e. the genome-wide SNP profiles and three subgroups (i.e. three canine breeds) in the population. Therefore, the number of dogs in each breed nor any information concerning potential breed-specific SNPs was used as input data. The accuracy of clustering these canine subjects according to the breeds was used as the major criterion for the evaluation of the robustness of the classification techniques developed. Additionally, the impacts of data processing, data cleanup and SNP annotation quality on the analysis result were also examined.

Of the three clusters generated, Cluster #1 (Table 2) closely resembled the group of Malinois, while Clusters #2 and #3 resembled the Labrador Retriever group (Table 3) and German Shepherd Dog group (Table 4), respectively. As shown in the Table 2, the algorithm developed can cluster the dogs of the Malinois breed (44 dogs) with accuracy >90%. Interestingly, the data process method, the annotation quality of the SNP, and the data cleanup method seemed to have only a minor effect on the accuracy of the clustering result. The result of Cluster #2 (Table 3) showed that this algorithm can categorize all Labrador Retriever dogs into one cluster with 100% accuracy. Consistent with the result of Cluster #1, the data process method, the annotation quality of the SNP and the data cleanup method had little effect on the accuracy of the clustering result. As with the Labrador Retriever clustering, this algorithm can accurately cluster the German Shepherd Dog group (47 dogs). Note that the accuracy of this result was close to 90%, and did not seem to be significantly affected by the data process method, the annotation quality of the SNPs or the data cleanup method.

**Table 2: Cluster 1 (45-47 Subjects).**

| Data File | Data Cleanup | Cluster 1 (44 Subjects) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | Correct | Incorrect | Missed | Accuracy (%) | Specificity (%) |
| DP Method 1_Full | No Cleanup | 45 | 42 | 3 | 2 | 95 | 93 |
| DP Method 1_Platinum | No Cleanup | ND | | | | | |
| DP Method 2_Full | No Cleanup | ND | | | | | |
| DP Method 2_Platinum | No Cleanup | 46 | 41 | 5 | 3 | 93 | 89 |
| DP Method 1_Full | Complete Cleanup* | 47 | 42 | 5 | 2 | 95 | 89 |
| DP Method 1_Platinum | Complete Cleanup* | 47 | 42 | 5 | 2 | 95 | 89 |
| DP Method 2_Full | Complete Cleanup* | 46 | 40 | 6 | 4 | 91 | 87 |
| DP Method 2_Platinum | Complete Cleanup* | 46 | 41 | 5 | 3 | 93 | 89 |

*The SNP will be excluded if there was a „no call" value.

**Table 3: Cluster 2 (26 Subjects).**

| Data File | Data Cleanup | Cluster 2 (26 Subjects) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | Correct | Incorrect | Missed | Accuracy (%) | Specificity (%) |
| DP Method 1_Full | No Cleanup | 26 | 26 | 0 | 0 | 100 | 100 |
| DP Method 1_Platinum | No Cleanup | ND | | | | | |
| DP Method 2_Full | No Cleanup | ND | | | | | |
| DP Method 2_Platinum | No Cleanup | 26 | 26 | 0 | 0 | 100 | 100 |
| DP Method 1_Full | Complete Cleanup* | 26 | 26 | 0 | 0 | 100 | 100 |
| DP Method 1_Platinum | Complete Cleanup* | 26 | 26 | 0 | 0 | 100 | 100 |
| DP Method 2_Full | Complete Cleanup* | 26 | 26 | 0 | 0 | 100 | 100 |
| DP Method 2_Platinum | Complete Cleanup* | 26 | 26 | 0 | 0 | 100 | 100 |

*The SNP will be excluded if there was a „no call" value.

**Table 4: Cluster 3 (47 Subjects).**

| Data File | Data Cleanup | Cluster 3 (47 Subjects) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | Correct | Incorrect | Missed | Accuracy (%) | Specificity (%) |
| DP Method 1_Full | No Cleanup | 46 | 44 | 2 | 3 | 94 | 96 |
| DP Method 1_Platinum | No Cleanup | ND | | | | | |
| DP Method 2_Full | No Cleanup | ND | | | | | |
| DP Method 2_Platinum | No Cleanup | 45 | 42 | 3 | 5 | 89 | 93 |
| DP Method 1_Full | Complete Cleanup* | 44 | 42 | 2 | 5 | 89 | 95 |
| DP Method 1_Platinum | Complete Cleanup* | 44 | 42 | 2 | 5 | 89 | 95 |
| DP Method 2_Full | Complete Cleanup* | 45 | 42 | 3 | 5 | 89 | 93 |
| DP Method 2_Platinum | Complete Cleanup* | 45 | 42 | 3 | 5 | 89 | 93 |

*The SNP will be excluded if there was a „no call" value.

## 4. CONCLUSIONS AND FUTURE DIRECTIONS

As shown in the results presented using this dataset of 117 canine subjects, we have developed a method that can provide an effective means to accurately classify subjects based on a common phenotype (in this case, the breed of the subjects) with no *a priori* criteria – a major advantage of this classification technique. Utilizing this method, SNP profiles can be evaluated solely based on their aggregated distance to achieve the overall clustering pattern. With further refinements (*e.g.* the incorporation of the Classification and Regression Trees (CART) methodology), this

method could potentially allow the identification of SNP predictors for different aspects of a phenotype/trait. As the CART algorithm can partition the data through a set of sequential binary splits based upon a single covariate at a time, it could be very useful in the genetic modeling of more complex traits that covariate information can be used to identify genetically homogeneous subgroups. In the case of complex traits like intelligence, in one intelligence behavior test, a SNP set (i.e. a synthesized feature) may be an important predictor, while in another test, different SNP set(s) may be critical.

## 5. REFERENCES


Almasy, L, Blangero, J. (2009) "Human QTL linkage mapping." *Genetica* **136**:333-340.

Amos, CI. (2007) "Successful design and conduct of genome-wide association studies." *Hum Mol Genet* **16**:220-225.

Borecki IB, Province MA. (2008) "Linkage and association: basic concepts." *Adv Genet* **60**:51-74.

Butcher, LM, Davis, OS, Craig, IW, Plomin, R. (2008) "Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays." *Genes Brain Behav*. **7**:435-446.

Ding, C, Jin S. (2009) "High-throughput methods for SNP genotyping." *Methods Mol Biol*. **578**:245-254.

Farrall, M, Morris, AP. (2005) "Gearing up for genome-wide gene-association studies." *Hum Mol Genet* **14**:157-162.

Gu CC, Todorov A, Rao DC. (1996) "Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci." *Genet Epidemiol* **13**:513-533.

Gu CC and Rao DC. (2003) "Designing an optimum genetic association study using dense SNP markers and family-based sample." *Front Biosci* **8**:s68-80.

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM. (2003) "Control of confounding of genetic associations in stratified populations." *Am J Hum Genet* **72**:1492-1504.

Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, Mendrick D, Tong W. (2010) "Evaluating variations of genotype calling: a potential source of spurious associations in genome-wide association studies." *J Genet* **89**:55-64.

Karlsson, EK, Baranowska, I, Wade, CM *et al*. (2007) "Efficient mapping of mendelian traits in dogs through genome-wide association." *Nat Genet* **39**:1321-1328.

Lander E, Kruglyak L. (1995) "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." *Nat Genet.* **11**:241-247.

Lindblad-Toh, K, Wade, CM, *et al.* (2005) "Genome sequence, comparative analysis and haplotype structure of the domestic dog." *Nature* **438**:803-819.

Miclaus K, Wolfinger R, Vega S, Chierici M, Furlanello C, Lambert C, Hong H, Zhang L, Yin S, Goodsaid F. (2010) "Batch effects in the BRLMM genotype calling algorithm influence GWAS results for the Affymetrix 500K array." *Pharmacogenomics J.* **10**:336-346.

Ostrander, EA, Galibert, F, Patterson, DF. (2000) "Canine genetics comes of age." *TIG* **16**:117-124.

Ostrander, EA, Wayne RK. (2005) "The canine genome." *Genome Res* **15**:1706-1716.

Pearson, TA, Manolio TA. (2008) "How to interpret a Genome-wide Association Study." *JAMA* **299**:1335-1344.

Sham, PC, Cherny SS, Purcell S. (2009) "Application of genome-wide SNP data for uncovering pairwise relationships and quantitative trait loci." *Genetica* **136**:237-243.

Ward, JH, Hook, ME. "A Hierarchical Grouping Procedure Applied to a Problem of Grouping Profiles." Lackland Air Force Base, Texas: Personal Laboratory, Wright Air Development Division, March 1961.

Wayne, RK, Ostrander, EA. (2007) "Lessons learned from the dog genome." *Trends in Genetics* **23**:557-567.

## 6. LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

BRLMM – Bayesian Robust Linear Model with Mahalanobis distance classifier
CART – classification and regression trees
CGP – co-evolutionary genetic programming
cM – centi Morgan
CITP – canine intelligence testing protocol
EDTA – ethylenediaminetetraacetic acid
GWAS – genome-wide association study
LD – linkage disequilibrium
MWD – military working dog
OD – optical density
PCR – polymerase chain reaction
PM – perfect match

QC – quality control
QTL – quantitative trait loci
SNP – single nucleotide polymorphism
TE – Tris + EDTA
TBE – Tris + Boric Acid + EDTA
WGSA – whole genome sampling assay